Article Watch, December 2012

Clive Slaughter

GHSU-UGA Medical Partnership, Athens, GA

This column highlights recently published articles that are of interest to the readership of this publication. We encourage ABRF members to forward information on articles they feel are important and useful to Clive Slaughter, GHSU–UGA Medical Partnership, UGA Health Sciences Campus, HSC Russell Hall – Room 164, 1425 Prince Ave, Athens, GA 30606, USA; phone: 706-713-2216; E-mail: cslaught@uga.edu; or to any member of the editorial board. Article summaries reflect the reviewer's opinions and not necessarily those of the association.

DNA SEQUENCING AND CHARACTERIZATION

Peters B A, Kermani B G, Sparks A B, Alferov O, Hong P, Alexeev A, Jiang Y, Dahl F, Tang Y T, Haas J, Robasky K, Zaranek A W, Lee J-H, Ball M P, Peterson J E, Perazich H, Yeung G, Liu J, Chen L, Kennemer M I, Pothuraju K, Konvicka K, Tsoupko-Sitnikov M, Pant K P, Ebert J C, Nilsen G B, Baccash J, Halpern A L, Church G M, Drmanac R. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* 2012;487: 190–195.

Sequencing methods that rely on short reads have proven efficient at generating large amounts of sequencing information quickly and cheaply but typically result in the loss of crucial phasing (haplotype) information for single nucleotide polymorphisms (SNPs) beyond a few kilobases. The present publication presents a streamlined, low-cost methodology for sequencing and haplotyping. The method relies on separating high molecular mass DNA fragments by dilution into the wells of a 384-well plate so that the chance of corresponding fragments from the maternal and paternal genomes ending up in the same well is reduced to ~5%. The genomic DNA is then amplified, fragmented, and ligated to unique barcode adapters in a series of steps performed within the same wells without intervening purifications. The contents of the 384 wells are combined, purified, and sequenced using the Complete Genomics platform. The process begins with just 100–130 pg DNA (the amount in 10–20 human cells) and achieves a redundancy of 19–38×, which ensures accuracy of variant calling and phasing. Accurate phasing of up to 97% of detected SNPs is demonstrated. Sequencing the maternal and paternal chromosomes in separate aliquots 10–40× reduces the error rate for single nucleotide variants to 1 in 10 megabases. The capabilities of this methodology promise to yield important information for comprehensive genetic studies, and the capabilities are crucial for clinical application of sequencing. Accurate screening of microbiopsies, circulating tumor cells, and preimplantation embryos generated by in vitro fertilization may be envisioned.

doi: 10.7171/jbt.12-2304-005

MACROMOLECULAR SYNTHESIS AND SYNTHETIC BIOLOGY

Wang J, Friedman G, Doyon Y, Wang N S, Li C J, Miller J C, Hua K L, Yan J J, Babiarz J E, Gregory P D, Holmes M C. Targeted gene addition to a predetermined site in the human genome using a ZFN-based nicking enzyme. *Genome Research* 22;2012:1316–1326.

Kim E, Kim S, Kim D H, Choi B-S, Choi I-Y, Kim J-S. Precision genome engineering with programmable DNA-nicking enzymes. *Genome Research* 2012;22:1327–1333.

Zinc-finger nucleases are customized enzymes designed to produce dsDNA breaks at predetermined sequence locations. Their zinc-finger domains bind to sequences flanking a target site. A sequence-nonspecific nuclease domain cleaves DNA at that site once subunits with appropriate specificity become dimerized on the target sequence. This approach to genome engineering harnesses the cellular processes that repair double-stranded breaks to accomplish targeted genomic alterations. However, the process of nonhomologous end-joining, which cells rely on to repair double-stranded breaks, is imprecise: it produces random, small insertions and deletions (indels) at the target site even in the presence of homologous donor DNA. Furthermore, the zinc-finger nucleases produce off-target mutations that make isolation of gene-edited cells difficult. Two groups have now re-engineered the nuclease

domain of zinc-finger nucleases to produce single-stranded rather than double-stranded breaks. Single-stranded breaks are repaired by the more precise process of homologous recombination, and off-target breaks are sealed faithfully. The site-specific "nickases" are expected to improve the use of genome engineering, especially in applications such as stem cell research and gene therapy, where precise editing is crucial.

CARBOHYDRATES AND GLYCOCONJUGATES

Walker S H, Carlisle B C, Muddiman D C. Systematic comparison of reverse phase and hydrophilic interaction liquid chromatography platforms for the analysis of N-linked glycans. *Analytical Chemistry* 84;2012:8198–8206.

Separation of N-linked glycans cleaved from proteins for the purpose of on-line mass spectrometric analysis is widely performed by hydrophilic interaction chromatography or porous, graphitized carbon chromatography. These modes of separation are compatible with the very hydrophilic glycans but are disadvantageous in a number of respects, including the rather poor resolution they provide. Reverse-phase chromatography, a technique capable of higher resolution, is unsuitable, as the hydrophilic glycans are poorly retained. This publication illustrates the advantages to be gained from hydrazine derivatization of glycans. The derivatization allows reverse-phase separation to be performed successfully. Peak capacities are increased, mass spectra are simplified by absence of ammonium derivatization, and column equilibration times are reduced. The derivatization takes only 4 h, and the derivatives are suitable for electrospray ionization mass analysis.

PROTEINS—PURIFICATION AND CHARACTERIZATION

Frank E G, McDonald J P, Karata K, Huston D, Woodgate R. A strategy for the expression of recombinant proteins traditionally hard to purify. *Analytical Biochemistry* 429; 2012:132–139.

The authors here express and purify proteins that have proven difficult to produce in soluble form in *Escherichia coli* by expressing the proteins at low, basal levels to facilitate their correct folding, rather than by overexpressing them in the conventional manner—an approach that results in the formation of insoluble inclusion bodies. Expression vectors are maintained here at the low level of approximately five copies/cell, and no external induction is performed. Under these circumstances, cell cultures can be conveniently harvested at any time, although the cell density at time of harvest may be chosen for greatest yield of soluble protein. Affinity tags are incorporated into the protein to facilitate purification in the normal way. The method is shown to enable the human DNA polymerases η, ι, and ν to be isolated from *E. coli* in active form. The origin of replication in vectors used here is compatible with the replicons of standard high-expression systems such as pET vectors (EMD Biosciences). When expressing multisubunit protein complexes, this feature allows nonproblematic components to be expressed at high levels, while the problematic component can be expressed at a low level within the same culture.

Boutet S, Lomb L, Williams G J, Barends T R M, Aquila A, Doak R B, Weierstall U, DePonte D P, Steinbrener J, Shoeman R L, Messerschmidt M, Barty A, White T A, Kassemeyer S, Kirian R A, Seibert M M, Montanez P A, Kenney C, Herbst R, Hart P, Pines J, Haller G, Gruner S M, Philipp H T, Tate M W, Hromalik M, Koerner L J, van Bakel N, Morse J, Ghonsalves W, Arnlund D, Bogan M J, Caleman C, Fromme R, Hampton C Y, Hunter M S, Johansson L C, Katona G, Kupitz C, Liang M, Martin A V, Nass K, Redecke L, Stellato F, Timneanu N, Wang D, Zatsepin N A, Schafer D, Defever J, Neutze R, Fromme P, Spence J C H, Chapman H N, Schlichting I. High-resolution protein structure determination by serial femtosecond crystallography. *Science* 2012;337:362–364.

Analysis of macromolecular structures by X-ray crystallography has traditionally demanded large crysals. The reason for this requirement is reduction of radiation damage to the crystal, which necessitates lowering the fluence of incident radiation. This, in turn, requires large crystals to achieve sufficient diffraction intensities, while reducing the radiation dose to individual molecules in the crytstal. Small crystals may be studied using microfocus synchrotron beamlines, but rapid damage to these small crystals remains a challenge. The present authors use X-ray-free electron laser radiation, which is sufficiently intense to yield high-quality diffraction patterns from radiation pulses of femtosecond duration. This duration is brief enough to avoid significant radiation damage. This practice effectively decouples the relationship among crystal size, radiation damage, and resolution in structural biology. The importance of this development is that microcrystals for many maromolecular complexes and membrane proteins are readily available, even though large crystals have proven difficult to obtain. With the use of the Coherent X-Ray Imaging instrument at the Stanford Linear Accelerator Center, the authors introduce a microjet of randomly oriented microcrystals of a model protein—hen egg white lysosyme—into the single-pulse beam. The crystals have dimensions of ~1 μm × 1 μm × 3 μm. In the beam, these crystals are subjected to X-ray pulses of 5- or 40-fs duration at a radiation dose that represents the classical limit for damage to cryogenically cooled crystals. The 40-fs data yield dif-

fraction patterns showing excellent statistics to 1.9 Å resolution. The results agree with synchrotron data. No evidence for radiation damage is evident. The methodology is anticipated to be immediately relevant to high-resolution studies of many macromolecules.

Klammt C, Maslennikov I, Bayrhuber M, Eichmann C, Vajpai N, Chiu E J C, Blain K Y, Esquivies L, Kwon J H J, Balana B, Pieper U, Sali A, Slesinger P A, Kwiatkowski W, Riek R, Choe S. Facile backbone structure determination of human membrane proteins by NMR spectroscopy. *Nature Methods* 2012;9:834–839.

Although as many as 30% of human genes encode integral membrane proteins, structures for only 30 human integral membrane proteins have become available in the Protein Data Bank. Recently, cell-free expression systems have been developed to overcome many of the limitations in expressing integral membrane proteins in bacterial hosts. In the present paper, these expression techniques are coupled with a combinatorial dual-labeling strategy that greatly facilitates the normally laborious sequential assignment of resonances from integral membrane proteins. The methodology also incorporates the use of an exogenous or covalently bound paramagnetic group that produces paramagnetic relaxation enhancement to provide constraints for long-range structure determination. The authors have used these strategies to deduce the backbone structures of six human integral membrane proteins within the brief period of 18 months. The technology is expected to accelerate acquisition of structural information for integral membrane proteins and hence to provide critical information for designing pharmaceuticals targeted against this class of proteins.

Babu M, Vlasblom J, Pu S, Guo X, Graham C, Bean B D M, Burston H E, Vizeacoumar F J, Snider J, Phanse S, Fong V, Tam Y Y C, Davey M, Hnatshak O, Bajaj N, Chandran S, Punna T, Christopolous C, Wong V, Yu A, Zhong G, Li J, Stagljar I, Conibear E, Wodak S J, Emili A, Greenblatt J F. Interaction landscape of membrane-protein complexes in *Saccharomyces cerevisiae*. *Nature* 2012;489:585–589.

An inventory of interactions among yeast membrane proteins (including 1144 integral, 400 peripheral, and 46 lipid-anchored membrane proteins) is presented in this report. Membrane proteins are purified by tandem affinity purification in the presence of nondenaturing detergents (Triton X-100, n-dodecyl-β-D-maltopyranoside, or octaethylene glycol monododecyl ether), which preserve protein interactions. Copurifying proteins are then identified by tandem mass spectrometry. The results specify 1726 high-confidence interactions and 501 heteromeric complexes. Novel complexes and novel components of known complexes are described. The data constitute a global map of the organization of yeast membranes in this prototypic eukaryotic cell type.

PROTEOMICS

Shuford C M, Sederoff R R, Chiang V L, Muddiman D C. Peptide production and decay rates affect the quantitative accuracy of protein cleavage isotope dilution mass spectrometry (PC-IDMS). *Molecular & Cellular Proteomics* 2012;11:814–823.

Proteins are commonly quantified after proteolytic digestion by introducing tracers that consist of known quantities of stable isotope-labeled peptides with sequences identical to peptides generated during digestion of the protein. In such a scheme, quantification relies on comparing signal strengths of the labeled peptide and the protein-derived peptide. The question is: when should the labeled peptides be added—before the proteolytic digestion begins? concurrently with addition of the protease? or after the digestion is complete? This study presents a comparison of these three options. The study notes discrepancies in quantification for nearly all of the peptides investigated depending on the time of addition. Some differences are as great as 30×. These results are explained in terms of the rates of production of the protein-derived peptides as digestion procedes and the rates of decay of these peptides and of the labeled standard. The authors suggest that adding labeled standards concurrently with the proteolytic enzyme is the method least likely to bias the results. Investigators using quantification methods of this kind are advised to be mindful of the biases that might arise from peptide instability and to place a high priority on choosing stable peptides for quantification standards.

FUNCTIONAL GENOMICS AND PROTEOMICS

The ENCODE Project Consortium, Bernstein B E, Birney E, Dunham I, Green E D, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57–74.

During the last 3 months, results from the Encyclopedia of DNA Elements (ENCODE) project have been published in an extensive series of papers in several journals. The present paper provides an overview of the work. The ENCODE project is of interest to readers of the *Journal of Biomolecular Techniques* both for the methodological advances the project has spawned and for the biological conclusions it has yielded. The aim of the ENCODE project is to describe all of the functional elements in the human genome. Examples of functional elements are elements that encode protein or RNA products, that bind proteins, that comprise recognizable chromatin signatures, that display evidence of evolutionary constraint, or that

display DNase hypersensitivity. The scale of the project is large. It has amassed 1640 datasets related to functional elements and has included study of 147 cell types.

A cardinal finding of the ENCODE project is that although much of the genome does not encode protein, it is definitely not junk. The majority (80%) of the genome contains elements that participate in biochemical functions. Intergenic regions are replete with enhancers (i.e., regulatory elements), promoters (i.e., sites involved with initiation of transcription), and sequences that encode untranslated RNA species. Approximately 75% of the genome is transcribed in at least one cell type. This finding provokes us to rethink the definition of the gene. Many SNPs associated with disease in genome-wide association studies lie within sequences that do not encode protein, and the majority resides in or close to functional elements.

The methodological advances achieved by the project include the development of computational procedures for documenting, validating and comparing data related to diverse functional elements. Among the fruits of these developments is the recognition that the distributions of DNA methylation, DNA accessibility, and transcription-factor binding sites are linked. For example, transcription-factor binding sites are less frequently methylated in cells that express the requisite transcription factors, suggesting that DNA methylation may be a passive process that occurs in regions that do not bind transcription factors.

Much work remains to be done. How do the functional elements regulate cellular processes through time? How different are the individual cells within tissues in the processes controlled by functional elements? How do the functional elements coordinate their functions within the network of interactions required to perform morphogenesis, establish phenotype, and develop specific diseases?

Ramskold D, Luo S, Wang Y-C, Li R, Deng Q, Faridani O R, Daniels G A, Khrebtukova I, Loring J F, Laurent L C, Schroth G P, Sandberg R. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology* 2012;30:777–782.

mRNA-Seq protocols are capable of sequencing and quantifying mRNAs with amounts some 1000× less than conventional RNA sequencing methods. The present adaptation of RNA-Seq for work with RNA from single cells is distinguished from previous implementations of the method in that it generates and amplifies only full-length cDNA. This feature enhances coverage of transcript sequences and enables alternative transcripts to be studied. Additionally, it permits assessment of the reproducibility of results with RNA dilution series and enables assessment of technical reproducibility when amounts of starting material are limiting. With just 1 ng total RNA (corresponding to 50–1200 cells), expression-level variability is <2× greater than with standard mRNA-Seq. With 10 pg total RNA (corresponding to single cells), the chief limitation is presently that transcripts expressed at the lower levels show stochastic loss, although this limitation is expected to lessen as the technology evolves. The new methodology is applied to a study of circulating tumor cells from a patient with melanoma. Six circulating tumor cells isolated from peripheral blood are compared with two primary melanocytes. Although closely related to the primary tumor cells, they show specific differences in transcription pattern. These differences constitute putative biomarkers for circulating melanoma cells. The methodology is anticipated to be useful for genome-wide transcriptional profiling of rare cells isolated by cell sorting or laser-capture microdissection.

Mullokandov G, Baccarini A, Ruzo A, Jayaprakash A D, Tung N, Israelow B, Evans M J, Sachidanandam R, Brown B D. High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. *Nature Methods* 2012;9:840–846.

microRNAs (miRNAs) are regulators of gene expression that work by binding to target mRNA species. This binding either inhibits mRNA translation or accelerates mRNA degradation. The human genome encodes >400 different miRNAs. The authors report methods for high-throughput assessment of miRNA function in cells. An expression vector library is constructed that contains a reporter gene to which binding sequences targeted by endogenous cell miRNAs are coupled. The expression levels of the mRNA products, measured by deep sequencing, provide an assay for the activity of the cognate miRNAs. The rate of target transcription is adjusted to physiologic levels. A target library designed for expression at supraphysiologic levels is also constructed here to provide decoys that sequester cognate miRNAs for loss-of-function screening. Unexpectedly, the authors show that >60% of the miRNAs have no detectable activity in terms of these assays, suggesting that the number of functional miRNAs may be appreciably smaller than the number of miRNA species expressed in cells. Only the miRNAs expressed at the highest levels appear to be functional. Moreover, some miRNAs expressed at relatively high levels exhibit only weak activity. These results are of importance to investigators interested in miRNA-based regulation of gene expression.

Fan H C, Gu W, Wang J, Blumenfeld Y J, El-Sayed Y Y, Quake S R. Non-invasive prenatal measurement of the fetal genome. *Nature* 2012;487:320–324.

Plasma from maternal peripheral blood contains cell-free DNA from both the mother and the fetus. The fraction of fetal DNA increases with fetal age and may attain levels as high as 50% of the total cell-free DNA. This paper demonstrates sequencing of the entire fetal genome by shotgun sequencing of prenatal maternal plasma DNA. The contributions of maternal and fetal DNA are deduced by an allele-counting approach. Genetic information from the father is not required. The inherited maternal haplotypes are deduced by determining the relative representation of the two alleles at heterozygous loci, and the inherited paternal haplotypes are identified by alleles that are different from the maternal ones at loci where the mother is heterozygous, and then imputing allelic identity at linked positions. Alternatively, the same counting principles can be applied directly to targeted deep sequencing of exome-enriched DNA from the maternal plasma. Although this work is of considerable methodological interest, its use in the clinical context will depend on developing capabilities for predicting phenotype from a genomic sequence that is presently lacking. Meanwhile, the ethical issues surrounding the acquisition of global genomic information pertaining to unborn children represent a serious concern that demands the attention of all persons.

CELL BIOLOGY AND TISSUE ENGINEERING

Powell A A, Talasaz A H, Zhang H, Coram M A, Reddy A, Deng G, Telli M L, Advani R H, Carlson R W, Mollick J A, Sheth S, Kurian A W, Ford J M, Stockdale F E, Quake S R, Pease R F, Mindrinos M N, Bhanot G, Dairkee S H, Davis R W, Jeffrey S S. Single cell profiling of circulating tumor cells: transcriptional heterogeneity and diversity from breast cancer cell lines. *PLoS ONE* 2012;7:e33788.

Understanding how tumor cells metastasize is a critical step in therapeutic control of the spread of cancer. Although rare circulating tumor cells in the blood of cancer patients have been identified, subjecting them to molecular analysis depends on achieving the exceptionally difficult task of isolating them from leukocytes that are present in vastly greater numbers. Powell et al. have surmounted this longstanding challenge by combining an efficient cell-separation procedure for enrichment of circulating tumor cells with single-cell transcriptional profiling for discovering the molecular basis of their phenotypic characteristics. For cell purification, they use an immunomagnetic enrichment device named a MagSweeper, which was developed in their own laboratory and licensed to Illumina. With the use of stringent criteria to identify circulating tumor cells, they isolate cells meeting the requisite criteria from 14/20 patients with primary breast cancer and from 21/30 patients with metastatic breast cancer. No such cells are captured from 20 patients with nonepithelial cancers or from 25 healthy subjects. The authors then conduct transcriptional profiling of single cancer cells using a microfluidic device in which to perform a RT-PCR assay. They profile the expression of 87 genes in 510 individual cells from 50 breast cancer patients. Most of the circulating tumor cells are triple-negative, whether or not the primary tumors of the patients display this phenotype. The triple-negative phenotype is associated with high metastatic potential and shorter time to metastasis. This result suggests an explanation for the failure of therapies targeting these biomarkers in some patients. The phenotypes of circulating cancer cells do not cluster by patient or by disease stage of the tumor tissue, indicating that studying the primary tumor alone may provide a suboptimal basis for making treatment decisions. The phenotypes of circulating cancer cells are distinct from those of tumor cell lines, suggesting that cell lines are not necessarily good models for tumors in vivo. The circulating cells show elevated expression levels of genes associated with metastasis and with epithelial-to-mesenchymal transition, indicating how expression levels help explain cell phenotypic characteristics. The methodology presented in this study is expected to contribute to biological understanding of metastasis and to clinical management of epithelial cancers.

IMAGING

Berk V, Fong J C N, Dempsey G T, Develioglu O N, Zhuang X, Liphardt J, Yildiz F H, Chu S. Molecular architecture and assembly principles of *Vibrio cholera*e biofilms. *Scienc*e 2012;337: 236–239.

The biofilm matrices in which some bacteria grow confer resistance to immune clearance and to antibiotic treatment that is of major clinical importance. However, little is known about the spatial organization of biofilms that confer the mechanical properties so important to the cells that construct them. *Vibrio cholera*e forms its biofilm from four components: a polysaccharide called *Vibrio* polysaccharide (VPS) and three matrix proteins, RbmA, RbmC, and Bap1, which contain carbohydrate-binding domains. Berk et al. here introduce tags into the genes encoding these three proteins that permit the proteins to bind antibodies labeled with distinct fluorophores that are supplied in the growth medium. VPS is visualized directly by binding to Cy3-labeled wheat-germ agglutinin. The authors use this labeling strategy to determine the spatial architecture of developing biofilms using four-color confocal imaging to compile movies of growing colonies of cells as they develop from a single founder cell. By imaging the WT cells and strains deficient in each of the proteins comprising the biofilm, they are able to assign functions to each protein: Bap1 helps the biofilm to adhere to surfaces,

RbmC and Bap1 encapsulate cell clusters, and RbmA participates in cell–cell adhesion. The development of a three-dimensional (3D) biofilm requires a series of specific, mutually interdependent steps of protein and VPS synthesis, secretion, capture, and cross-linking. As the spatial resolution of confocal laser-scanning microscopy is insufficient to visualize the developmental sequence directly, the authors construct a multicolor 3D super-resolution imaging apparatus using stochastic optical resolution microscopy. This reveals that VPS is organized into 50- to 200-nm diameter spheroids that protrude from the cell surface. The results help explain how the biofilm matrix is capable of reforming, stretching, and expanding to accommodate cell growth as colonies develop.